Continual optimization and refactoring of GSI for Rapid Refresh Forecast System (RRFS): Current Status and Prospect from the EMC perspective

NATIONAL WEATHER SERVICE

July, 2023

Presenter: Ting Lei
Lynker Contractor at EMC/NCEP/NWS/NOAA
Ting Lei[1], Shun Liu[2], Eric Rogers[2], Xiaoyan Zhang[3], Miodrag Rancic[1], Manuel Pondeca[1], Matthew Pyle[2], Jacob Carley[2] ,Daryl Kleist[2]

A special acknowledgment goes to Wanshu Wu, the key developer of the initial GSI interface for the regional Fv3 among many other achievements, who had retired from EMC.

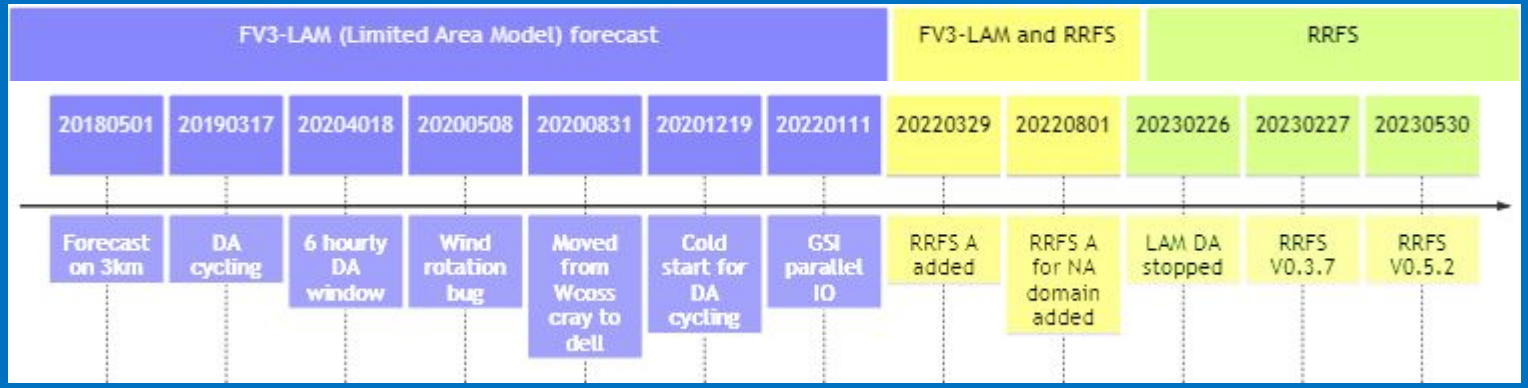
[1]Lynker at NOAA/NWS/NCEP/EMC
[2]NOAA/NWS/NCEP/EMC
[3]SAIC at NOAA/NWS/NCEP/EMC

Outline

- Background and Goal
- RFFS GSI computational scalability analysis
- Optimization works: past,ongoing and vision
- Summary and discussion

Background: RRFS and its' predecessor at EMC

FV3-LAM forecast with DA cycling: running quasi-operationally at EMC since 2019–a pilot of RRFS



The Regional Data assimilation team at EMC had first diligently developed the interface between the Finite-Volume Cubed-Sphere (FV3) Limited Area Model (LAM) and the Gridpoint Statistical Interpolation (GSI) and has been dedicated to its ongoing improvements, encompassing data assimilation methods, running flexibilities, and computational efficiency. We deeply value collaboration within our community and employ GitHub-based code sharing, which significantly enhances our efforts.

RRFS : Session 7 - Model Development and Application: RRFS Development and Planning

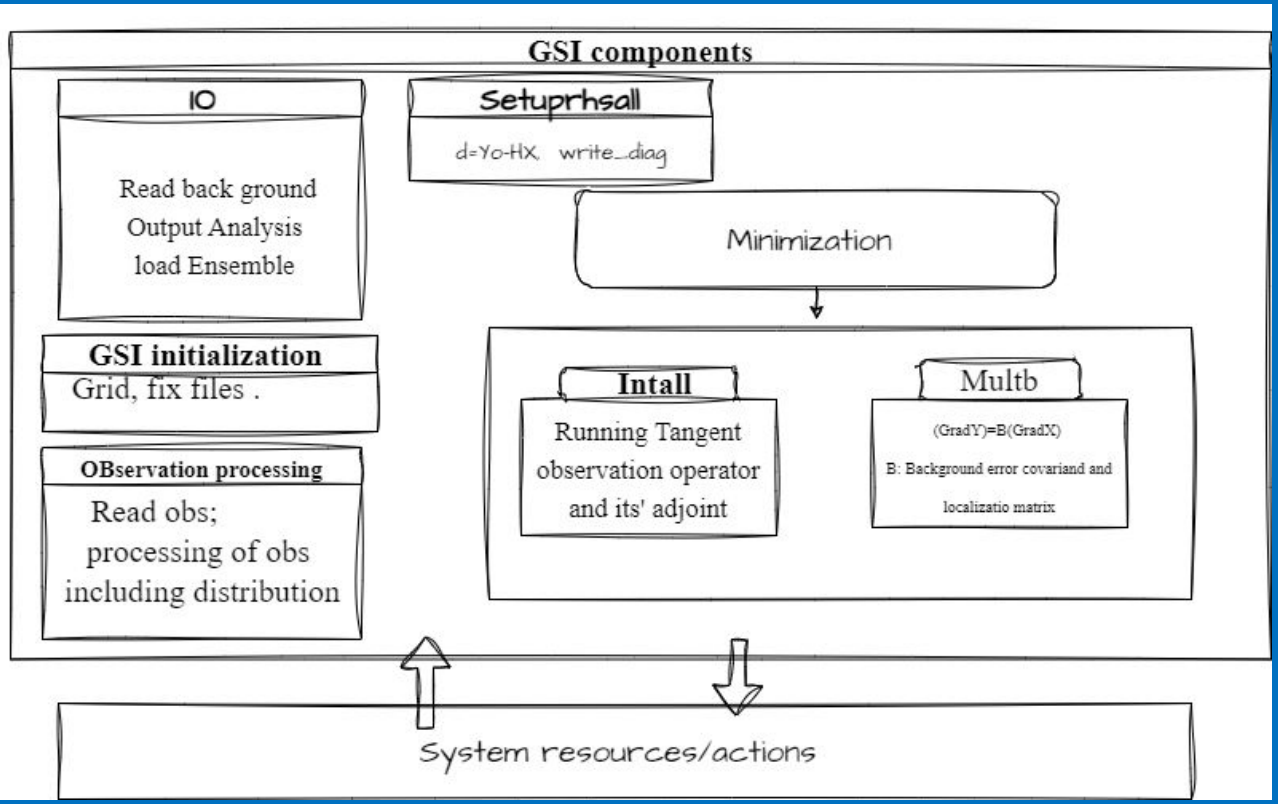
GOAL : Optimization of RRFS GSI at EMC

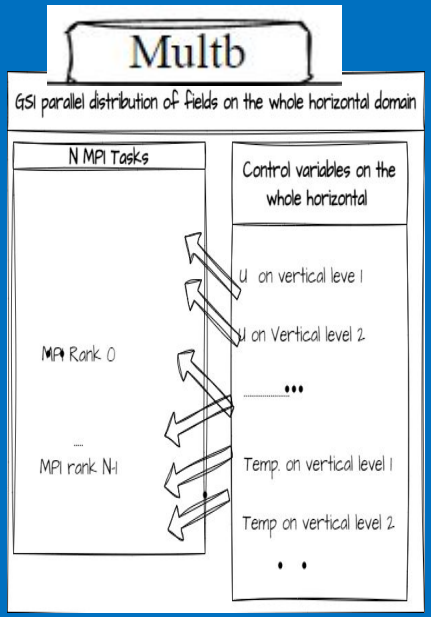
The optimization strategies discussed in this presentation are primarily focused on the utilization of MPI and OpenMP to enhance the scalability of the GSI variational system. For a more comprehensive understanding of EMC's work in this direction, it's beneficial to consider this in conjunction with the work documented by J. Derber on EMC GSI Github issues

For EnKF, An important upgrading by J. Whitaker was to let EnKF keep working with arrays sizes smaller than 2GB for extremely large setup as the North American RRFS. The optimization on EnKF of GSI at EMC is mainly on the IO for FV3-LAM, which will not be covered in this presentation.

Investigation and exploration of JEDI's computational performance has been initiated and some preliminary results in comparison with GSI is going to be described.

Computational scalability: key components in GSI





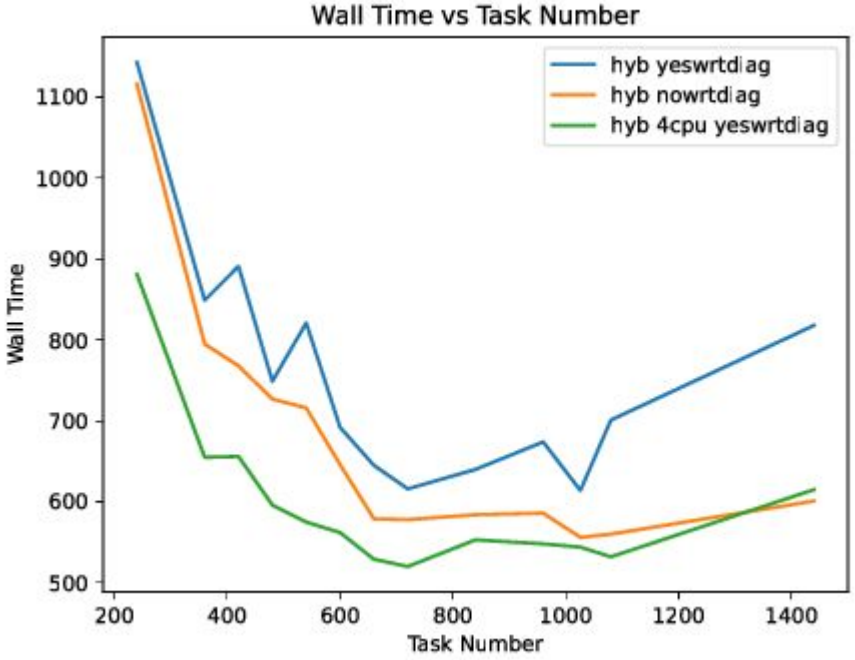
FV3-LAM IO optimization

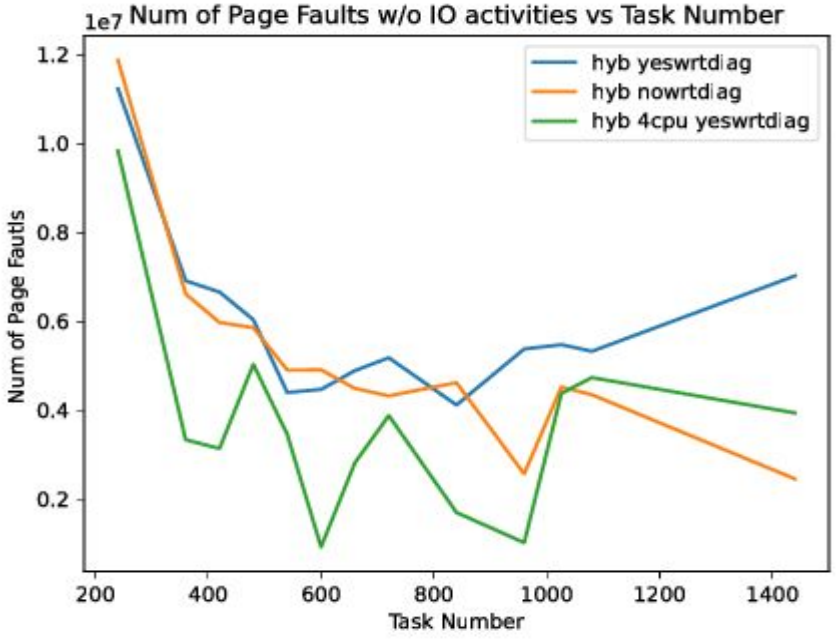
Computational scalability 1

Experiments with 3 Km conus RRFS Runs.
Grid: 1820X1092X65 ; four 3d and one 2d control variables.
30 FV3-LAM ensemble members, 015 for static B.
All experiments run on Hera of NOAA RDHPCS
No Openmp function is launched (omp_num_threads=1) for clarification

Experiment/setup	Number of processors for a MPI task	Write_diag
Hyb yeswrtdiag	2	Yes
Hyb nowrtdiag	2	No
Hyb 4 cpu yeswrtdiag	4	yes

Computational scalability 2

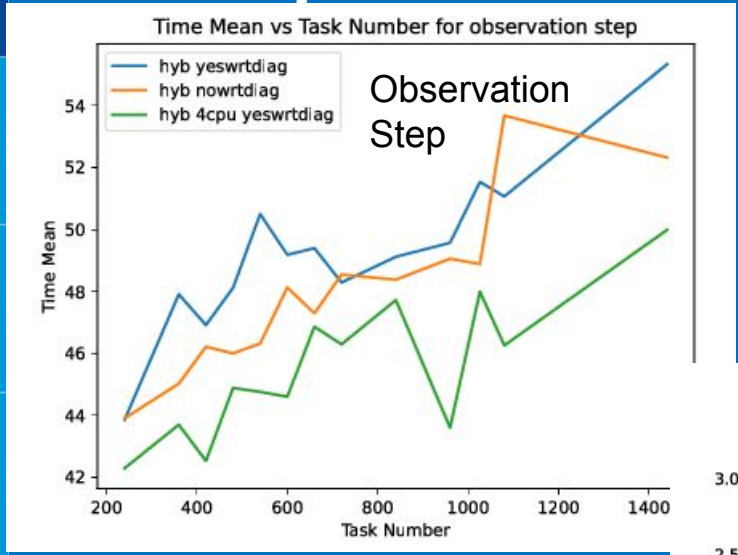


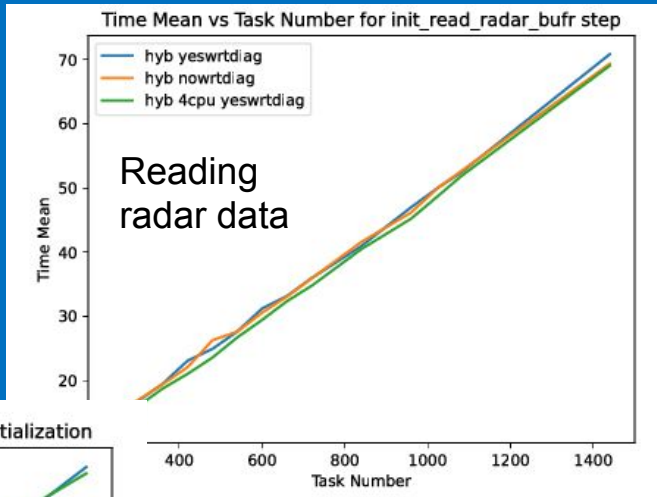


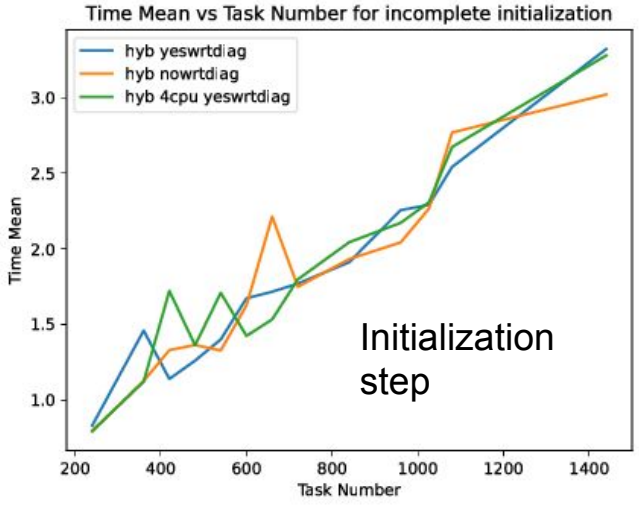
The “saturation point” number of MPI tasks is around 700 for 3km conus RRFS and expected to be larger for 3km North american RRFS runs.

The number of page faults without IO activities refer to number of switches between primary and secondary memory. it is also affected by the real time system status. In combination with following findings, it is regarded as the main cause for the differences among the 3 experiments.

Computational scalability 3



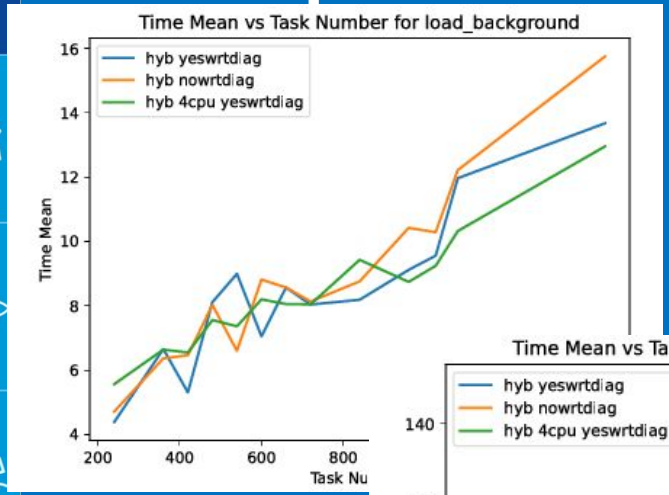


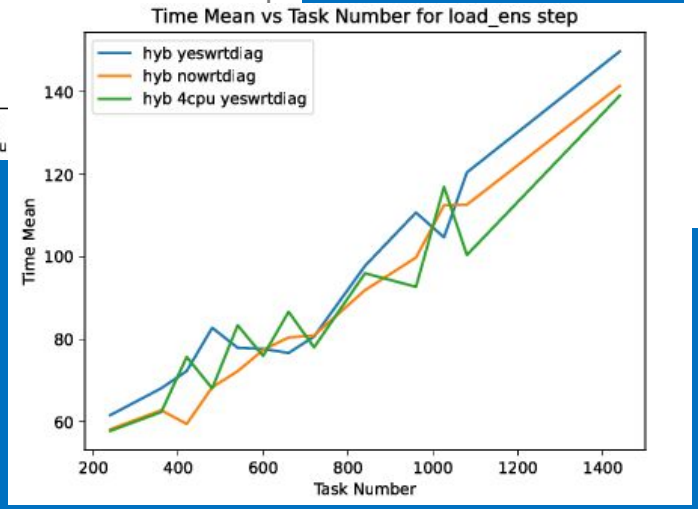


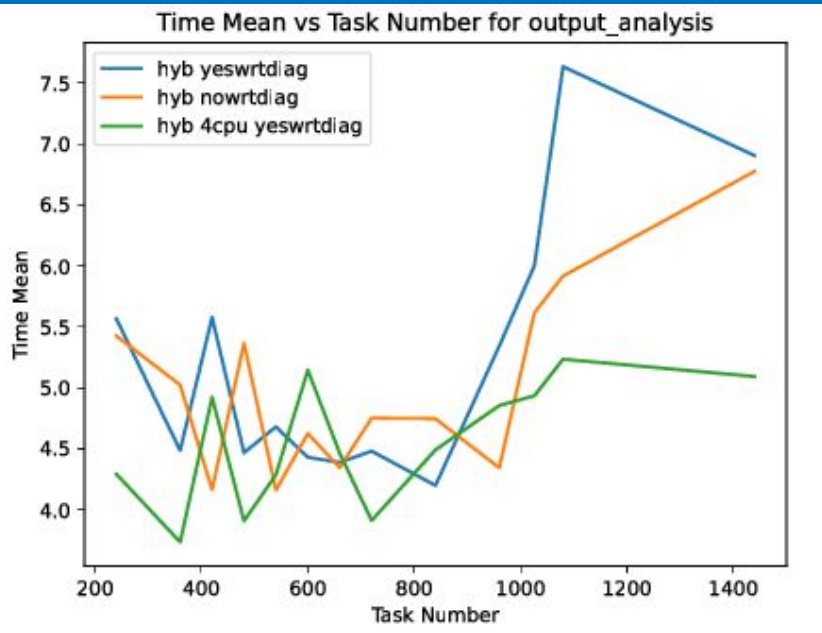
As expected, the observation step doesn't show good scalability but its cost is still tolerable.
The same for the GSI initialization step except for the init_read_radar step

init_read_rada should be revisited with other options like reading pre-calculated data

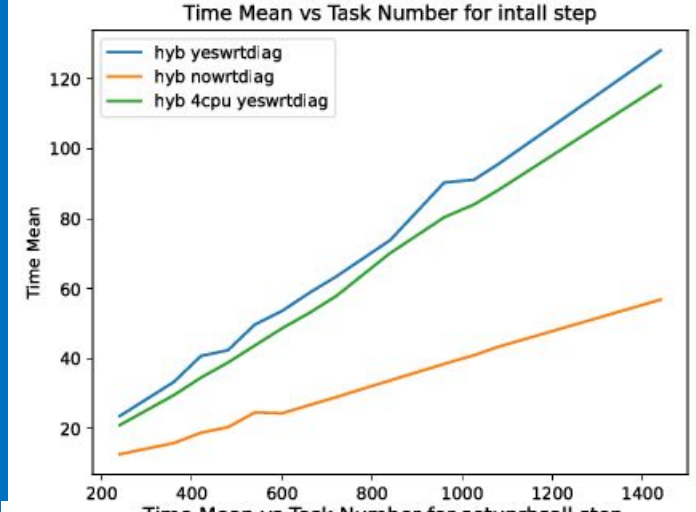
Computational scalability

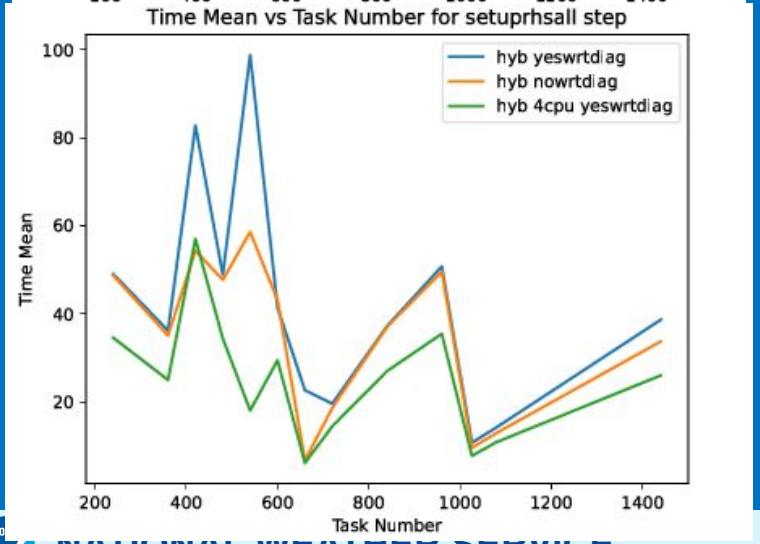


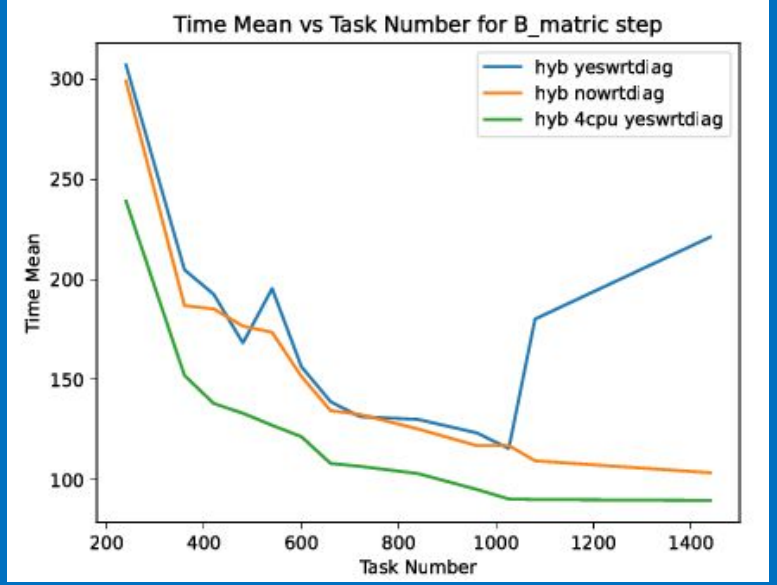




Output shows better computation performance/scalability
Loading of ensembles uses significant time and more and more time mainly for more reading time of a single member.
Enhanced parallelization should help.





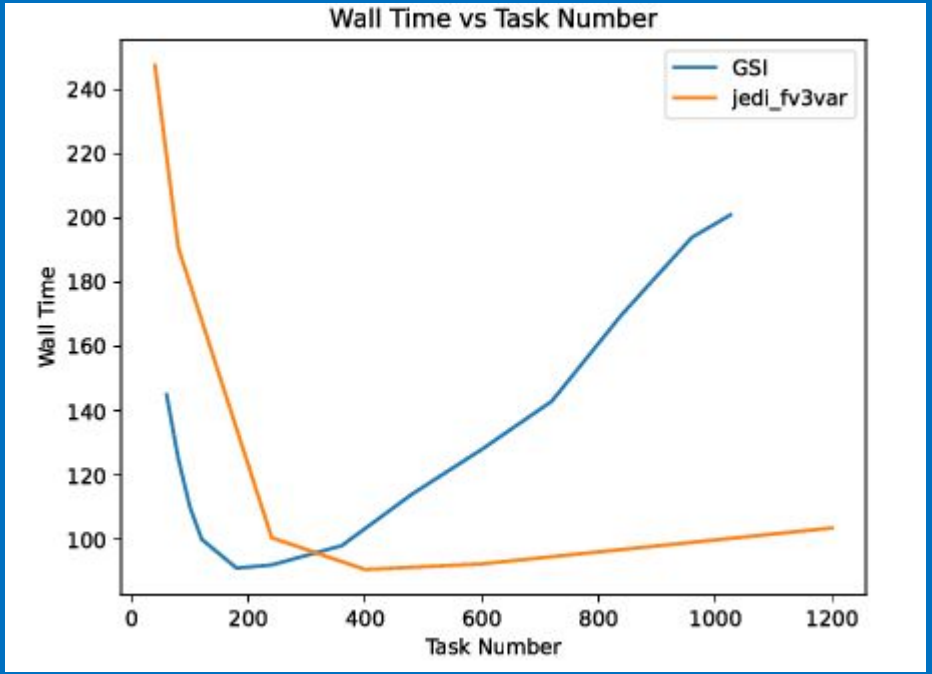


- After around 700 MPI tasks, the increase of mpi tasks has marginal impacts (theoretically it would still help vertical recursive filtering).
- The increased time for intall is already comparable to mutllib and need to be addressed.
- Setuprhsall showed smaller dependency on the task numbers and obvious vulnerabilities to runtime system performance.

JEDI Computational scalability

13k conus domain/grid :396X233X60
Pure Ens3DVar (Lei, etal.,2021).
(Cautions needed)

- To achieve their respective minimum clock times, which are strikingly similar, GSI uses only half the number of MPI tasks as JEDI.
- JEDI variational app shows much better scalability for larger MPI task numbers



GSI/JEDI optimization :

Parallel IO for fv3-lam, further parallelization for load_ens

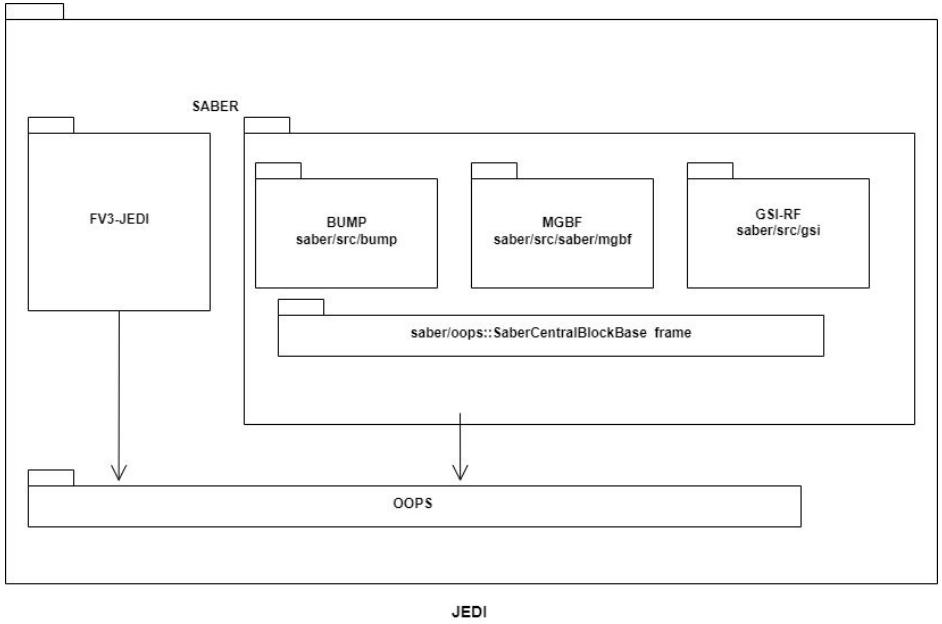
Multigrid Beta Function approach (J. Purser.et al.2022)

- GSI: M. Rancic's previous presentation (17.4) and more
- Jedi : Ongoing

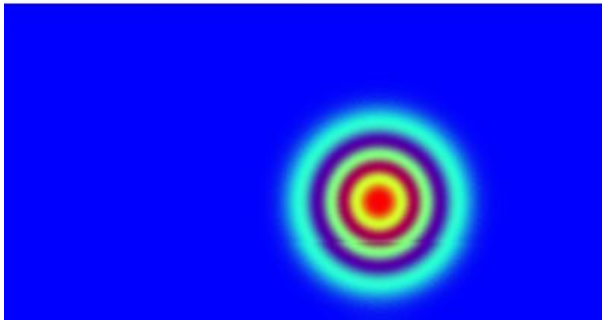
Further investigations on intall and setuprhsall steps while the imbalance from homogenous observation distributions is a general issue to variational DA.

Systematic computation scalability analysis for various JEDI application (including EnKF) at various setup.

MGBF in JEDI



A preliminary version of MGBF at SABER/JEDI has been implemented.
A Dirac test is shown below.



A generalized standing alone MGBF is being refactored/upgraded following OOPS principles, which is going to be applied to upgrade the mgbf component in SABER/JEDI and GSI

T.Lei thanks B. Menetrier at JCSDA/IRIT for his help in updating of MGBF to the refactored SaberCentralBlockBase and D. Holdaway for his GSI-RF of SABER as a good example in the design of the blueprint for MGBF SABER/JEDI

• Summary and discussion

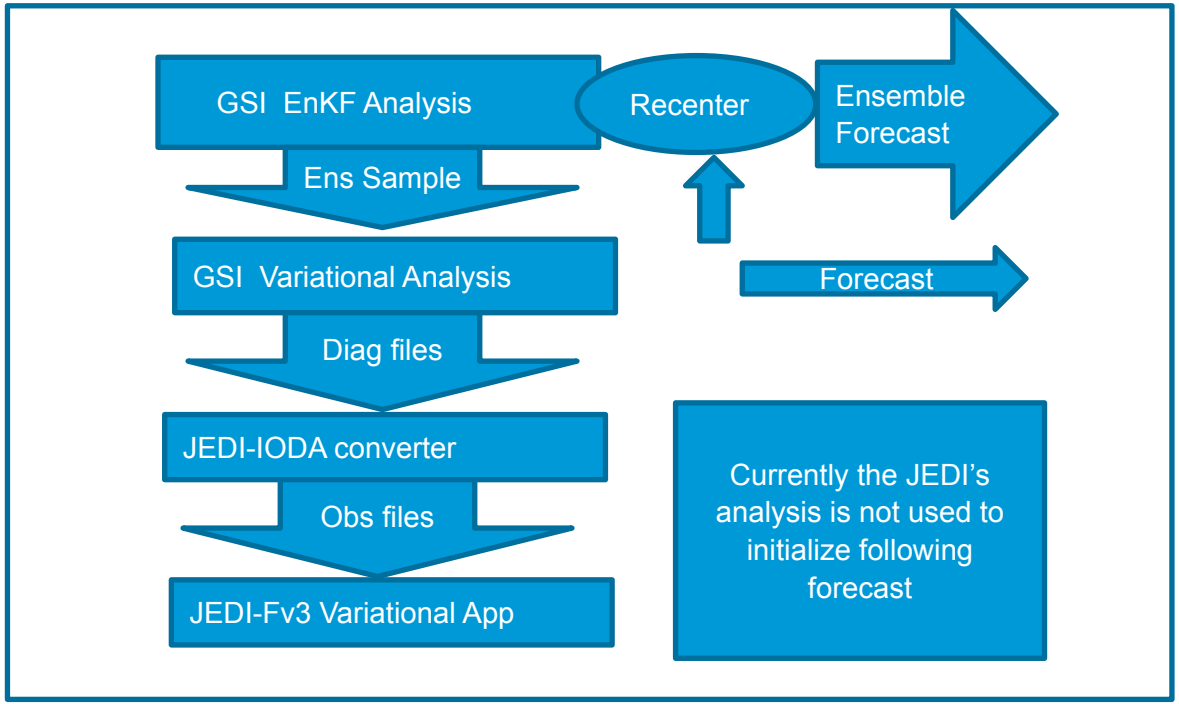
- The recursive filter based background error covariance modeling and the step invoking tangent linear observation model and its' adjoint are demonstrated to be major bottlenecks for GSI (the variational component) run with mpi task numbers of thousand orders.
- MGBF is proven to be powerful for the background matrix modeling process and provide a promising tool for JEDI.
- The impact of GSI 's "write diag " mainly through higher requirements for system memory on GSI performance are demonstrated and corresponding re-organization of the DA workflow are recommended.
- The inhomogeneous distributions of observations and related operations on MPI tasks in the current variation data assimilation is an issue .
- Preliminary comparison of GSI and JEDI variational app indicated their respective strength and weakness, while JEDI variational app demonstrates better scalability for larger MPI task numbers.
- Further investigation on the interaction between DA programs with the HPC systems in collaboration with HPC system engineers.

Reference

T. Lei,S. Liu,J. Carley,C. Martin,D. Holdaway, B. M´en´etrier,E. Rogers,X. Zhang,B. Blake,M. Hu,and D. Kleist, 2021: Tests of hybrid EnKF-Variational Data Assimilation capabilities using JEDI with NOAA's Next Generation Regional High Resolution NWP System, WCRP-WWWRP symposium, 2021, Virtual.

Purser, R. J., M. Rancic, and M. S. De Pondeca, 2022: The Multigrid Beta Function Approach for Modeling of Background Error Covariance in the Real Time Mesoscale Analysis (RTMA). Mon. Wea. Rev., Vol. 150, N. 4, 715-732.

The interface for JEDI variational app.



- Prescreening of observation files is added
- The observation operator is set to VertInterp for all types rather than the Identity operator for surface obs as the vertical coordinate convention is not the same for FV3 as for other models. A fix has recently been made by JEDI developers.
- Use the QC flag from GSI by using Domain filter for GsiUseFlag